

Praktyczne uczenie maszynowe w języku R

Fred Nwanganga
Mike Chapple

Przekład: Natalia Chounlamany-Turalska

Praktyczne uczenie maszynowe w języku R

First published in English under the title Practical Machine Learning in R, by Fred Nwanganga and Mike Chapple

Copyright © 2020 by John Wiley & Sons, Inc., Indianapolis, Indiana

This edition has been translated and published under licence from John Wiley & Sons, Inc.

John Wiley & Sons, Inc., takes no responsibility and shall not be made liable for the accuracy of the translation.

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage retrieval system, without permission from publisher.

Polish language edition published by APN PROMISE S.A., Copyright © 2022

Autoryzowany przekład z wydania w języku angielskim, zatytułowanego: Practical Machine Learning in R, by Fred Nwanganga and Mike Chapple, opublikowanego przez John Wiley & Sons, Inc.

Wszystkie prawa zastrzeżone. Żadna część niniejszej książki nie może być powielana ani rozpowszechniana w jakiegokolwiek formie i w jakikolwiek sposób (elektroniczny, mechaniczny), włącznie z fotokopiowaniem, nagrywaniem na taśmy lub przy użyciu innych systemów bez pisemnej zgody wydawcy.

APN PROMISE SA, ul. Domaniewska 44a, 02-672 Warszawa
tel. +48 22 35 51 600, fax +48 22 35 51 699
e-mail: wydawnictwo@promise.pl

Książka ta przedstawia poglądy i opinie autorów. Przykłady firm, produktów, osób i wydarzeń opisane w niniejszej książce są fikcyjne i nie odnoszą się do żadnych konkretnych firm, produktów, osób i wydarzeń, chyba że zostanie jednoznacznie stwierdzone, że jest inaczej. Ewentualne podobieństwo do jakiegokolwiek rzeczywistej firmy, organizacji, produktu, nazwy domeny, adresu poczty elektronicznej, logo, osoby, miejsca lub zdarzenia jest przypadkowe i niezamierzone.

Wszystkie znaki towarowe występujące w książce mogą być własnością ich odnośnych właścicieli.

APN PROMISE SA dołożyła wszelkich starań, aby zapewnić najwyższą jakość tej publikacji. Jednakże nikomu nie udziela się rękojmi ani gwarancji.

APN PROMISE SA nie jest w żadnym wypadku odpowiedzialna za jakiegokolwiek szkody będące następstwem korzystania z informacji zawartych w niniejszej publikacji, nawet jeśli APN PROMISE została powiadomiona o możliwości wystąpienia szkód.

ISBN: 978-83-7541-478-3 (druk), 978-83-7541-479-0 (ebook)

Przekład: Natalia Chounlamany-Turalska
Redakcja: Marek Włodarz
Korekta: Ewa Swędrowska
Skład i łamanie: MAWart Marek Włodarz

*Moim rodzicom Grace i Friday. Bez Was nie byłbym tym, kim jestem.
Dziękuję, że zawsze byliście przy mnie. Tęsknię za Wami.*

Wasz kochający syn Chuka

Ricky, jestem dumny z tego, jakim młodym mężczyzną się stałeś.

Kocham Cię, Tato

Spis treści

O autorach.....	xiii
O redaktorach technicznych.....	xiv
Podziękowania.....	xv
Wprowadzenie.....	xvii

Część 1 **Zaczynamy**

Rozdział 1	Czym jest uczenie maszynowe?.....	3
	Odkrywanie wiedzy w danych.....	4
	Wprowadzenie do algorytmów.....	4
	Sztuczna inteligencja, uczenie maszynowe i głębokie uczenie.....	5
	Techniki uczenia maszynowego.....	6
	Uczenie nadzorowane.....	7
	Uczenie nienadzorowane.....	11
	Wybór modelu.....	13
	Techniki klasyfikacji.....	13
	Techniki regresji.....	14
	Techniki uczenia się relacji podobieństwa.....	15
	Ocenianie modelu.....	15
	Błędy klasyfikacji.....	16
	Błędy regresji.....	18
	Typy błędów.....	19
	Partycjonowanie zbiorów danych.....	21
	Metoda wydzielenia.....	22
	Metody walidacji krzyżowej.....	22
	Ćwiczenia.....	23
Rozdział 2	Wprowadzenie do języka R i RStudio.....	25
	Witamy w świecie R.....	25
	Komponenty języka R i RStudio.....	27
	Język R.....	27
	RStudio.....	28
	RStudio Desktop.....	29
	RStudio Server.....	30
	Poznawanie środowiska RStudio.....	31
	Pakiety R.....	39

	Repozytorium CRAN	39
	Instalowanie pakietów	40
	Ładowanie pakietów	41
	Dokumentacja pakietu	42
	Pisanie i uruchamianie skryptu R	43
	Typy danych w języku R	46
	Wektory	47
	Sprawdzanie typów danych	49
	Przekształcanie typów danych	52
	Brakujące wartości	53
	Ćwiczenia	54
Rozdział 3	Zarządzanie danymi	55
	Tidyverse	55
	Zbieranie danych	56
	Kluczowe aspekty	57
	Zbieranie „prawdziwych” danych	57
	Relevantność danych	57
	Ilość danych	57
	Etyka	58
	Importowanie danych	58
	Wczytywanie plików CSV	58
	Wczytywanie innych plików rozdzielanych	61
	Eksploracja danych	62
	Opisywanie danych	62
	Wystąpienie	62
	Cecha	62
	Wymiarowość	64
	Rzadkość i gęstość	64
	Rozdzielczość	64
	Statystyki opisowe	64
	Wizualizowanie danych	71
	Porównanie	71
	Relacje	73
	Rozkład	74
	Skład	75
	Przygotowywanie danych	76
	Oczyszczanie danych	77
	Brakujące wartości	77
	Szum	81
	Wartości odstające	84

Nierównowaga klas	84
Przekształcanie danych	86
Normalizacja	86
Dyskretyzacja	91
Kodowanie zerojedynkowe	91
Redukcja danych	94
Próbkowanie	95
Redukcja wymiarowości	101
Ćwiczenia	102

Część 2 Regresja

Rozdział 4	Regresja liniowa	105
	Wypożyczanie rowerów a regresja	106
	Zależności między zmiennymi	107
	Korelacja	108
	Regresja	115
	Prosta regresja liniowa	116
	Zwykła metoda najmniejszych kwadratów	117
	Model prostej regresji liniowej	120
	Ocenianie modelu	120
	Reszty	121
	Współczynniki	122
	Dane diagnostyczne	122
	Wielokrotna regresja liniowa	124
	Model wielokrotnej regresji liniowej	125
	Ocenianie modelu	126
	Testy diagnostyczne reszt	127
	Analiza punktów wpływowych	130
	Współliniowość	134
	Ulepszanie modelu	135
	Uwzględnienie relacji nieliniowych	136
	Uwzględnienie zmiennych kategoryalnych	138
	Uwzględnienie interakcji między zmiennymi	141
	Wybieranie ważnych zmiennych	142
	Mocne i słabe strony	147
	Studium przypadku: przewidywanie ciśnienia krwi	148
	Importowanie danych	149
	Eksploracja danych	150
	Dopasowywanie modelu prostej regresji liniowej	152

	Dopasowywanie modelu wielokrotnej regresji liniowej	153
	Ćwiczenia.	162
Rozdział 5	Regresja logistyczna	165
	Poszukiwanie potencjalnych darczyńców	166
	Klasyfikacja	168
	Regresja logistyczna.	169
	Iloraz szans.	171
	Dwumianowy model regresji logistycznej	174
	Rozwiązywanie problemu brakujących wartości	176
	Rozwiązywanie problemu danych odstających	180
	Podział danych	185
	Rozwiązywanie problemu nierównowagi klas	186
	Trenowanie modelu	188
	Ocenianie modelu.	188
	Współczynniki	191
	Dane diagnostyczne	193
	Dokładność predykcji.	194
	Ulepszanie modelu	196
	Rozwiązywanie problemu współliniowości	196
	Wybór wartości granicznej	203
	Mocne i słabe strony.	204
	Studium przypadku: przewidywanie dochodu	205
	Importowanie danych	206
	Eksploracja i przygotowywanie danych	206
	Trenowanie modelu.	210
	Ocenianie modelu.	213
	Ćwiczenia.	215

Część 3 Klasyfikacja

Rozdział 6	K najbliższych sąsiadów	221
	Wykrywanie choroby serca	222
	K najbliższych sąsiadów.	224
	Znajdowanie najbliższych sąsiadów	225
	Oznaczanie danych bez etykiet	228
	Wybieranie odpowiedniej wartości k	229
	Model k najbliższych sąsiadów	230
	Rozwiązywanie problemu brakujących danych	231
	Normalizowanie danych.	232
	Rozwiązywanie problemu cech kategoryalnych	233

	Dzielenie danych	235
	Klasyfikowanie nieoznaczonych danych	235
	Ocenianie modelu.	236
	Ulepszanie modelu	237
	Mocne i słabe strony.	239
	Studium przypadku: powrót do zbioru danych darczyńców	239
	Importowanie danych	240
	Eksploracja i przygotowywanie danych	240
	Rozwiązywanie problemu brakujących wartości.	241
	Normalizowanie danych.	243
	Dzielenie i równoważenie danych	245
	Budowanie modelu.	246
	Ocenianie modelu.	247
	Ćwiczenia.	248
Rozdział 7	Naiwny klasyfikator Bayesa	249
	Klasyfikowanie emaili jako spamu.	250
	Naiwna metoda Bayesa	251
	Prawdopodobieństwo	252
	Prawdopodobieństwo łączne	253
	Prawdopodobieństwo warunkowe	254
	Naiwna klasyfikacja Bayesa.	255
	Wygładzanie addytywne	259
	Naiwny model Bayesa	261
	Dzielenie danych	265
	Trenowanie modelu	265
	Ocenianie modelu.	266
	Mocne i słabe strony naiwnego klasyfikatora Bayesa	267
	Studium przypadku: powrót do problemu wykrywania choroby serca	268
	Importowanie danych	268
	Eksploracja i przygotowywanie danych	269
	Budowanie modelu.	271
	Ocenianie modelu.	272
	Ćwiczenia.	273
Rozdział 8	Drzewa decyzyjne	275
	Przewidywanie decyzji o pozwoleniu na budowę	276
	Drzewa decyzyjne	277
	Partycjonowanie rekurencyjne.	279
	Entropia	283
	Zysk informacyjny	284

Nieczystość Giniego	287
Przycinanie	287
Budowanie modelu drzewa klasyfikacyjnego	289
Podział danych	292
Trenowanie modelu	293
Ocenianie modelu	293
Mocne i słabe strony modelu drzewa decyzyjnego	296
Studium przypadku: powrót do problemu przewidywania dochodu ...	297
Importowanie danych	298
Eksploracja i przygotowywanie danych	298
Budowanie modelu	300
Ocenianie modelu	300
Ćwiczenia	302

Część 4 Szacowanie i podnoszenie wydajności

Rozdział 9	Ocenianie wydajności	305
	Szacowanie przyszłej wydajności	305
	Walidacja krzyżowa	308
	<i>k</i> -krotna walidacja krzyżowa	308
	Walidacja krzyżowa Leave-One-Out	312
	Losowa walidacja krzyżowa	314
	Próbkowanie metodą bootstrap	316
	Poza dokładnością predykcji	318
	Kappa	320
	Precyzja i kompletność	323
	Czułość i swoistość	326
	Wizualizacja wydajności modelu	329
	Krzywa ROC	330
	Obszar pod krzywą	334
	Ćwiczenia	336
Rozdział 10	Ulepszanie wydajności	339
	Dostrajanie parametrów	339
	Automatyczne dostrajanie parametrów	340
	Niestandardowe dostrajanie parametrów	345
	Metody zespołowe	351
	Bagging	352
	Boosting	356
	Stacking	359
	Ćwiczenia	364

Część 5 **Uczenie nienadzorowane**

Rozdział 11	Odkrywanie wzorców za pomocą reguł asocjacyjnych	367
	Analiza koszykowa	368
	Reguły asocjacyjne	368
	Identyfikowanie silnych reguł	370
	Wsparcie	370
	Ufność	371
	Przyrost	371
	Algorytm Apriori	372
	Odkrywanie reguł asocjacyjnych	374
	Generowanie reguł	375
	Ocenianie reguł	380
	Mocne i słabe strony	384
	Studium przypadku: identyfikowanie wzorców zakupów spożywczych	384
	Importowanie danych	385
	Eksploracja i przygotowywanie danych	385
	Generowanie reguł	387
	Ocenianie reguł	387
	Ćwiczenia	390
	Uwagi	391
Rozdział 12	Grupowanie danych poprzez klasteryzację	393
	Klasteryzacja	393
	Klasteryzacja metodą k średnich	397
	Segmentowanie uczelni poprzez klasteryzację metodą k średnich	400
	Tworzenie klastrów	401
	Analizowanie klastrów	404
	Wybieranie odpowiedniej liczby klastrów	406
	Metoda „łokcia”	406
	Metoda średniego zarysu	408
	Statystyka odstępu	409
	Mocne i słabe strony klasteryzacji metodą k średnich	411
	Studium przypadku: segmentowanie klientów galerii handlowej	412
	Eksploracja i przygotowywanie danych	412
	Klasteryzacja danych	413
	Ocenianie klastrów	415
	Uwagi	416
	Ćwiczenia	416
	Indeks	417